第2章

PCI Expressデバイス&システム設計の基礎の基礎



——プロトコルやデバイスの機能を理解する

野崎原生

ここでは，PCI Express対応デバイスおよびシステムを設計するために最低限知っておくべき知識を解説する．前半で，シリアル通信全体に当てはまる技術を説明する．後半は，PCI Expressシステムを構成するデバイスの働きやアーキテクチャ，プロトコルについて重要なポイントをまとめる．（編集部）

PCI Expressは，従来のPCIやPCI-Xとソフトウェア上の互換性を保ったまま，ハードウェアについてパラレル・バスからシリアル通信へ変更したものです．

では，なぜパラレルからシリアルへ変更しなければならなかったのでしょうか．また，それによって何が改善されたのでしょうか．PCI Expressの説明に入る前に，これらについて少し復習しましょう．

1. なぜパラレルからシリアルへ移行するのか

パラレル・バスからシリアル通信へ変更する理由としては，次のようなクロック遅延の問題が挙げられます．

1.1 コモン・クロックとソース・シンクロナス

コモン・クロック方式は，PCIやPCI-X 1.0，そのほか多くのパラレル・バスで用いられています．名まえのとおりデータの送信側と受信側が，共通のクロックに同期してデータ転送を行うものです．転送を正しく行うためには，送信側から受信側までの遅延が1クロック周期以内に収まっている必要があります．逆に言えば，「クロック周期を遅延時間より短くできない」，イコール「転送速度を上げられない」ということになります．

式(1)はPCIにおけるタイミング・バジェット（見積もり）ですが，遅延時間（とセットアップ，クロック・スキューの合計）がクロック周期以下になっていなければならないことを示しています（**図1**）．

$$T_{val} + T_{prop} + T_{su} + T_{skew} \leqq T_{cyc} \quad \cdots\cdots(1)$$

T_{val}：クロックの立ち上がりからデータ出力までの遅延時間，T_{prop}：ボード上の伝播遅延時間，T_{su}：受信側のセットアップ時間，T_{skew}：送信側と受信側のクロック・スキュー，T_{cyc}：クロック周期

一方，ソース・シンクロナス方式はPCI-X 2.0，DDR SDRAMなどで用いられています．コモン・クロックの問

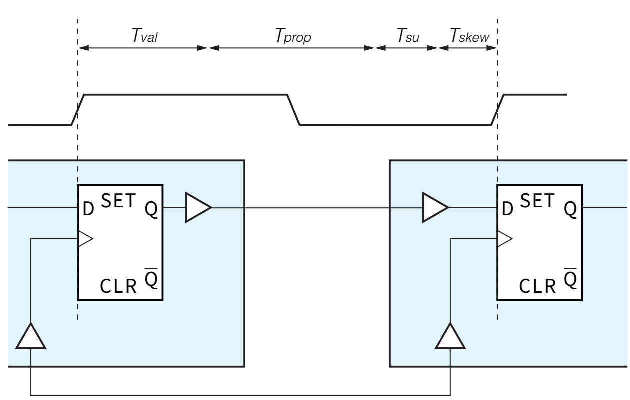


図1　コモン・クロック方式のタイミング・バジェット
PCIに用いられるコモン・クロック方式では，遅延時間がクロック周期以下になっていなければならない．

題を解決するため，データとそれをラッチするストローブを同一デバイスからドライブします．データとそのラッチ信号がほぼ同じ経路を通って受信側へ伝わるため，伝播遅延が相殺されてタイミング・バジェット（見積もり）において考慮する必要がなくなります．

式(2)にPCI-X 2.0のタイミング・バジェットを示します．PCIでは存在したT_{val}やT_{prop}がなくなり，遅延時間よりも短い周期でデータを転送できることがわかります（**図2**）．

$$\left.\begin{array}{l} T_{tdvfh}+T_{ssskew}\leqq T_{idvfh} \\ T_{tfhdx}+T_{ssskew}\leqq T_{ifhdx} \end{array}\right\} \quad \cdots\cdots(2)$$

T_{tdvfh}：受信側のセットアップ時間，T_{tfhdx}：受信側のホールド時間，T_{ssskew}：データ信号間およびそれらとストローブの間のスキュー，T_{idvfh}：送信側でのストローブ前のデータ有効期間，T_{ifhdx}：送信側でのストローブ後のデータ有効期間

しかし，さらに転送速度が上がると，データとストローブの間のスキューが問題になり，スキューの値（の2倍）より転送間隔を短くできなくなります．ボード製造時の誤差を特別に制御しないかぎりスキューは300ps～500ps程度なので，1Gbps前後がソース・シンクロナス方式の限界になります．

1.2 スキューの問題を解決するシリアル転送

シリアル転送方式を使うことによって，上記のスキューの問題を解決できます．シリアル転送はデータ信号が1本

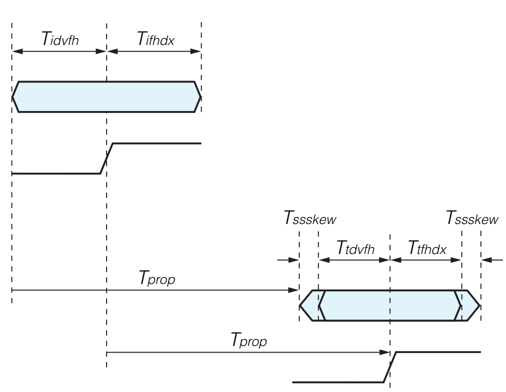


図2　ソース・シンクロナス方式のタイミング・バジェット
ソース・シンクロナス方式では，コモン・クロック方式の場合に存在したT_{val}やT_{prop}がなくなり，遅延時間よりも短い周期でデータを転送できる．

しかないため，データ間のスキューはもともと存在しません．また，データとクロック（ストローブ）を同じ信号で送るエンベデッド・クロックという符号化を行うことによって，データとストローブの間のスキューもなくなります．これにより，スキューによる制約を超えた速度でデータ転送を行えるようになります．

なお，PCI Expressは複数レーン（x4やx16）に対応していますが，これはパラレル・バスに戻っているわけではありません．一つ一つがエンベデッド・クロックを使ったシリアル・バスになっており，それを複数束ねることによって帯域幅を増やしているのです．

以下に，シリアル転送の特徴的な技術をまとめます．

- **エンベデッド・クロック**――エンベデッド・クロックは，その名のとおりデータにクロック情報を埋め込んで送信する方式．受信側で受け取ったビット列からデータとクロックを分離する．エンベデッド・クロックにはいくつかの符号化方式があるが，PCI Expressでは8b/10bと呼ばれるものを用いる．これは，Fibre Channelやギガビット Ethernetで使われているものと同じ方式である．
- **パケット通信**――パラレル転送と異なり，シリアル転送には専用の制御信号がないので，転送の開始と終了，つまりデータの切れ目を示す信号がない．その代わりに特殊なビット・パターンをデータの前後に付加して転送するパケット通信となる．PCI Expressでは，この前後に付加する特殊なビット・パターンにKコードを使っている．
- **差動伝送方式**――差動伝送方式は，「コモン・モード・ノイズに強い」，「差動電位はシングルエンド電位の2倍になるため，信号振幅を小さくしてもノイズ・マージンを確保しやすい」，「信号振幅を小さくすると同じスルーレートでも立ち上がり時間や立ち下がり時間が短くなる」という特徴があり，高速な信号転送には必要不可欠な技術である．差動伝送方式はシリアル転送特有の技術ではなく，HyperTransportのようなパラレル・バスでも用いられている．シリアル転送ではもともと信号本数が少ないため，差動にしても全体の配線がそれほど増えないという利点がある．

2. PCI Express システムの構成

図3に，一般的なPCI Express システムの構成例を示します．

2.1 通信単位――レーン，リンク，ポート

各コンポーネントは，ほかのコンポーネントと接続するためのポートを一つ以上持っています．通常，エンドポイントやブリッジは一つのポートだけを持ち，ルート・コンプレックスやスイッチは複数のポートを持っています．

ポートには，「レーン」と呼ばれるPCI Expressの基本となる通信単位が一つ以上含まれます．レーンを一つだけ含むポートは「x1ポート」と呼ばれます．二つ含むものは「x2ポート」と呼ばれ，それ以上は同様にx4/x8/x12/x16/x32ポートまで対応できます．一般に，x1/x4/x8/x16がよく使われています．

レーンは，送信ペア（Tx＋/Tx－）と受信ペア（Rx＋/Rx－）の4本の信号から構成され，それぞれの方向で2.5 Gbpsの転送を行います．また，「Gen2」と呼ばれる第2世代のPCI Expressでは，5Gbpsへ転送速度を引き上げるための規格を現在（2005年11月）検討中です．

通信を行っている二つのコンポーネントの各ポート，それに含まれるレーンをまとめたものを「リンク」と呼びます．通常，ポートのレーン数とリンクのレーン数は同じですが，レーン数の異なるポートどうしをつなげた場合は一致しません．

2.2 システムを構成する四つのデバイス

システムを構成するデバイスとして，「ルート・コンプレックス」，「エンドポイント」，「スイッチ」，「ブリッジ」が挙げられます．

- **ルート・コンプレックス**――ルート・コンプレックスとは，PCI階層の根幹（root）となるデバイス．通常は複数のPCI Expressポートを持ち，各ポートがルートとなる．また，図3に示したように一般的にはCPUやメイン・メモリも接続される．
- **エンドポイント**――グラフィックス・コントローラやUSBホスト・コントローラのように，I/Oデバイスや周辺デバイスと呼ばれているものは，PCI Expressでは「エンドポイント」と呼ぶ．エンドポイントには，レガシ・エンドポイント，PCI Expressエンドポイント，ルート・コンプレックス内蔵エンドポイントの3種類がある．レガシ・エンドポイントは，従来製品との互換性を保つためにPCI Expressエンドポイントの要件を満たせない

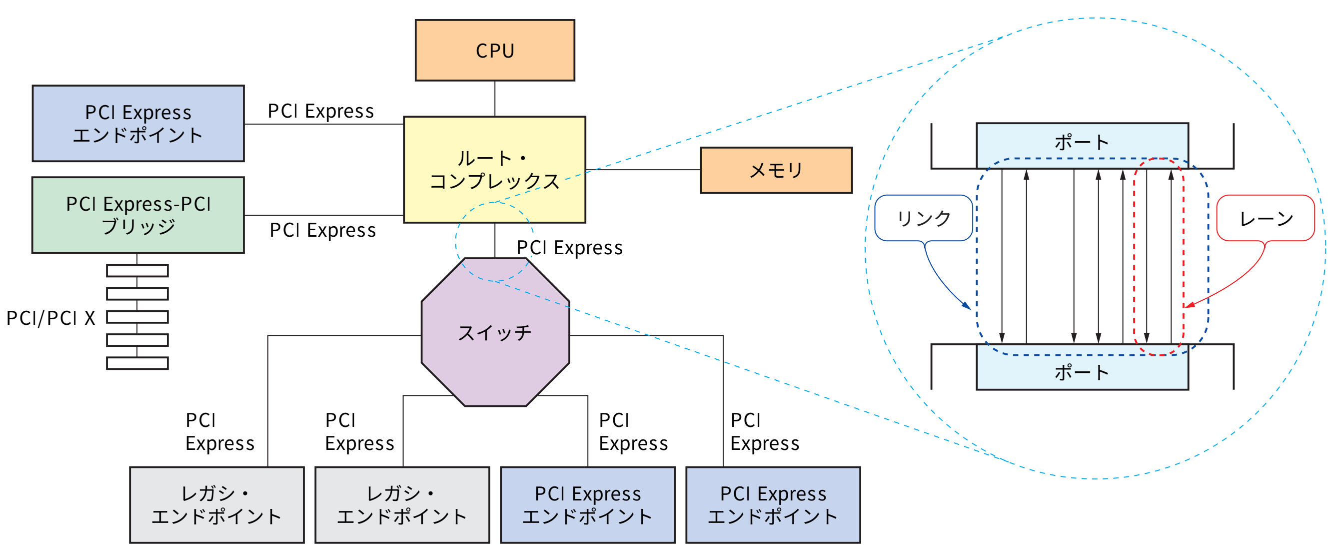


図3　PCI Express システム構成例
システムを構成するデバイスとして，「ルート・コンプレックス」，「エンドポイント」，「スイッチ」，「ブリッジ」が挙げられる．

もので，I/Oアクセスやロック・サイクルを使うことが許されている．PCI Expressエンドポイントは，一般的な周辺デバイスであり，「タイプ0のコンフィグレーション空間を持つ」，「I/Oサイクルを使わない」，「64ビット・アドレスに対応する」，「MSIまたはMSI-Xをサポートする（下掲のコラムを参照）」といった要件を満たす必要がある．ルート・コンプレックス内蔵エンドポイントは，その名のとおりルート・コンプレックスに内蔵されたタイプ0のコンフィグレーション空間を持つエンドポイント（例えば，メモリ・コントローラなど）．

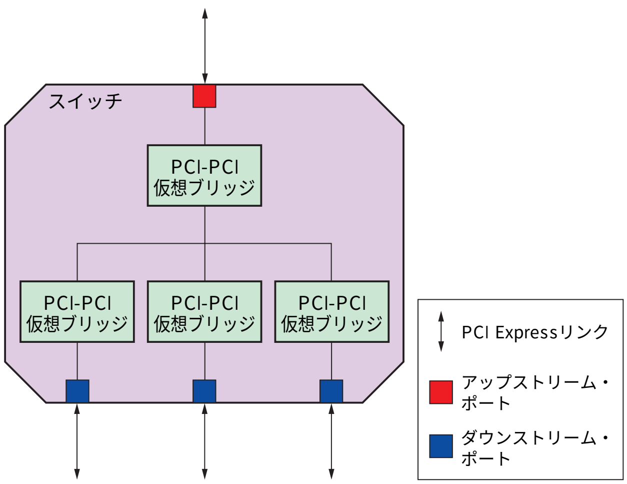


図4　スイッチの構造
スイッチにはアップストリーム・ポートと呼ばれるルート・コンプレックスに近い側のポートが一つと，ダウンストリーム・ポートと呼ばれるルート・コンプレックスから遠い側のポートが複数含まれる．

- **スイッチ**――スイッチは，PCI Expressポートのファンアウトを増やすために用いられるデバイス．ソフトウェアからは，図4のように複数のPCI-PCIブリッジに見える．スイッチにはアップストリーム・ポートと呼ばれるルート・コンプレックスに近い側のポートが一つと，ダウンストリーム・ポートと呼ばれるルート・コンプレックスから遠い側のポートが複数含まれる．
- **ブリッジ**――ブリッジとは，PCI ExpressシステムにPCIやPCI-Xのデバイスやカードを接続するためにプロトコル変換を行うデバイス．ソフトウェアからはPCI-PCIブリッジに見える．

● クロックICと電源ICの選びかた

PCI Expressシステムでは，以上の主要なコンポーネントのほかに，クロックICや電源ICなどの部品も重要な役割を果たしています．

クロックICは，各PCI Expressデバイスへリファレンス・クロックを供給しています．PCI Expressリンクを形成するデバイス間の周波数偏差は±300ppm以内（1ppmは百万分の1の誤差）と決められていますが，従来のPCIやPCI-Xのように，各デバイス間のクロック・スキューを考慮する必要がないため等長配線は不要です．

また，±300ppmの周波数偏差さえ満たしていれば，同じクロック源でなくてもかまいません．しかし，オプショ

Column　PCI Expressの電力管理，割り込み，エラー処理

ここでは，パワー・マネージメントや割り込みなどについて，従来のPCIとPCI Expressの相違点をまとめます．

● ソフトウェアを介さず自動的に電力を抑える仕様を追加

PCI Expressでは，従来のソフトウェアによって制御されるPCIパワー・マネージメントに加えて，デバイスがアイドル状態を検出した場合にソフトウェアの介在なしに自動的にリンクをパワー・セーブ状態に遷移させるASPM（Active State Power Management）が追加されました．

ASPMで用いられるリンクのパワー・セーブ状態として，L0sとL1の二つの状態が定義されています．なお，L0sとは，パワーはそれほど下がらないがL0への復帰が早い（数十ns～数μs）状態です．また，L1とは，パワーは下がるがL0への復帰に時間がかかる（数μs～数十μs）状態です．

● 割り込みにはMSIを使用する

PCIやPCI-Xでは，INTA#，INTB#，INTC#，INTD#という4本の信号（サイドバンド信号）を使って割り込みを通知していました．

PCI Expressではサイドバンド信号がなくなり，割り込みはメッセージを使って送信される方式に変更されました．レベル割り込みであるINTx#をエミュレートするため，Assert_INTxとDeassert_INTxという2種類のメッセージが用意されています．割り込みをアサートしたいデバイスは，システムに対してAssert_INTxメッセージを送信します．割り込みが処理されてほかに割り込み要求がなくなると，Deassert_INTxメッセージを送ります．ただし，この方式は従来のソフトウェアとの互換性を保つためのもので，PCI Expressとしては推奨されない割り込み方式です．

PCIではオプションだったMSI（Message Signaled Interrupt）は，PCI Expressでは必須となります．こちらが推奨の割り込み方式

ンのSSC（spread spectrum clocking）を使う場合でもリンクの両端の周波数偏差は±300ppm以内でなければならないため，通常は同じクロック源からのリファレンス・クロックを使います．

電源ICについては，PCI Expressだからといって特別な電源が必要になるわけではありません．しかし，ホットプラグに対応する場合はスロットごとに電源をON/OFFする必要があるため，通常は専用の電源ICを使って実現します．

3. PCI Expressのプロトコル

PCI Expressのプロトコルは，トランザクション層，データリンク層，物理層から構成されます（図5）．このような階層構造をとることで，仕様を拡張しやすくしています．例えば，Gen2でデータ転送速度を5Gbpsとしても，物理層だけが変更され，ほかの層を変える必要はありません．

3.1 トランザクション層——上位から下位へ渡すTLPの構築

トランザクション層は，PCI Expressプロトコルの最上位層に位置します．デバイス・コアとデータの受け渡しを行い，そのデータを下位層と受け渡すために以下の処理を行います．

- TLP（Transaction Layer Packet）の構築と処理
- フロー制御
- 仮想チャネルの管理
- トランザクション・オーダリングの処理

● メッセージを伝達するための専用パケットTLP

TLPはヘッダから始まり，（必要であれば）データ・ペイロード，オプションのTLPダイジェストまたはECRC（End-to-end CRC）が続きます（図6）．ヘッダの長さはトランザクションの種類によって3DW（double word；1

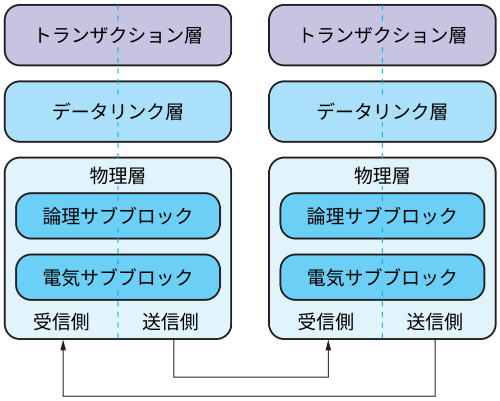


図5 プロトコル階層
PCI Expressのプロトコルは，トランザクション層，データリンク層，物理層から構成される．

2

です．

● デバイスはホットプラグへの対応が必須

PCIやPCI-Xではホットプラグはオプションのため，仕様も別になっていました．一方，PCI Expressデバイスは，かならずホットプラグをサポートしなければなりません．そのため，ホットプラグ制御用レジスタも本文の3.4項で述べた“PCI Express Capability Registers（機能レジスタ）”に含まれています．

なお，PCI Expressシステムがホットプラグをサポートするかどうかはオプションです．実際には，サーバやワークステーションのハイエンド機種のみがホットプラグをサポートするのではないかと予想されています．

● エラー処理の種類は三つに

PCIやPCI-Xではパリティ・エラーとシステム・エラーの2種類のエラーしかなく，かつエラーの定義はデバイス任せでした．

PCI Expressでは，コレクタブル・エラー，ノンフェイタル・エラー，フェイタル・エラーの3種類に増えて，どのようなエラーがどのエラー・クラスに分類されるのかも厳密に定義されました．標準の機能では，「これら三つのうちのいずれかのエラーが発生した」ということしかわかりません．しかし，オプションのアドバンスト・エラー・リポートに対応しているデバイスでは，さらに細かいエラーの発生要因やエラーが起きたときのTLPのログなどがレジスタに記録され，詳細にエラーの原因を解析できます．

なお，これらのエラーも，サイドバンド信号ではなくメッセージを使ってシステムへ通知されます．

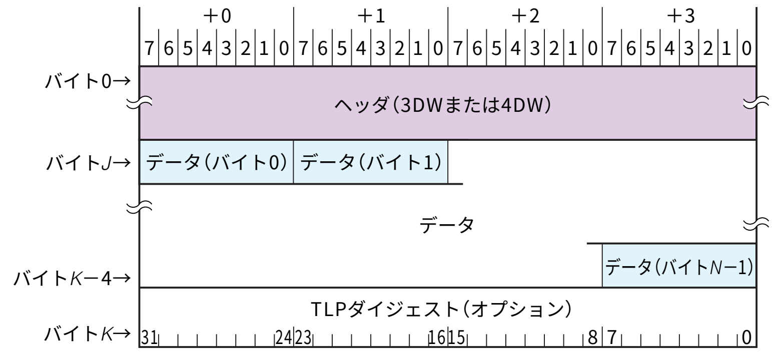


図6
TLPの構造
TLPはヘッダから始まり，(必要であれば)データ・ペイロード，オプションのTLPダイジェスト(ECRC)が続く．

表1
TLPの分類

TLPの種類	ヘッダ長	データ・ペイロード
メモリ・リード・リクエスト	3DW(32ビット・アドレス) 4DW(64ビット・アドレス)	なし
メモリ・ライト・リクエスト	3DW(32ビット・アドレス) 4DW(64ビット・アドレス)	あり
I/Oリード・リクエスト	3DW	なし
I/Oライト・リクエスト	3DW	あり
コンフィグレーション・リード・リクエスト	3DW	なし
コンフィグレーション・ライト・リクエスト	3DW	あり
コンプリーション	3DW	なし(ライト) あり(リード)
メッセージ	4DW	なし/あり

DW＝32ビットとする)または4DWとなります．プロトコルとしては，データ・ペイロードは0バイト～4,096バイトに設定可能ですが，実際の最大値はデバイスごとに決まっています．そのため，まずはじめにソフトウェアがすべてのデバイスのレジスタ(Max_Payload_Size_Supported)を読み，どのデバイスでもサポートできるサイズに制限するようにレジスタ(Max_Payload_Size)を設定します．データの後にはオプションで1DWのTLPダイジェストを付加することができます．

TLPは，トランザクションの種類や方向によって**表1**のように分類されます．

メッセージは，PCI Expressで新しく追加されたトランザクションです．従来のPCIでは，割り込み，PME(Power Management Event)，エラーなどの送信にサイドバンド信号を使っていましたが，PCI Expressではメッセージを用います(pp.34-35のコラム「PCI Expressの電力管理，割り込み，エラー処理」を参照)．また，上記以外にも新しいTLPとしてセキュリティを強固にするためのTrusted Configurationや，I/Oの仮想化に使われるアドレス変換などが現在検討されています．

TLPヘッダの始めの1DWはすべてのトランザクションで共通の仕様になっていて，ヘッダ長，データ・ペイロードのあり/なしを示すフォーマット(Fmt)，トランザクション種別，データ・ペイロード長などからなります．ヘッダの2DW以降はトランザクションごとに仕様が決まっています(**図7**)．

● 受信バッファの空きを確認して送受信を行うフロー制御

PCI Expressではクレジット・ベースのフロー制御，すなわち受信側に十分な空きがあるかどうかをチェックしてから送信を行います．リセット後にリンクが確立したとき，デバイスどうしが自分のバッファ・サイズをお互いに通知することによって，リンク先の受信バッファのサイズを知ります．また，受信側は受け取ったデータの処理が終わってバッファに空きができると，それを送信側へ通知します．なお，これらの通知は実際にはデータリンク層が行います．

こうしたフロー制御はPCI Expressのオーダリング・ルールを守るため，ポステッド・リクエスト，ノンポステッド・リクエスト，コンプリーションで独立に制御されます．さらに，それぞれがヘッダとデータ・ペイロードで別々に制御されます．その結果，PCI Expressでは，PH(Posted Header)，PD(Posted Data)，NPH(Non Posted Header)，

R：予約
Fmt：フォーマット
TC：トラフィック・クラス
TD：TLPダイジェスト
（ECRCが付与されていることを示す）
EP：デバイス内部のバッファでパリティ・エラーなどが発生してデータ品質に問題が発生した場合，意図的にCRCエラーとなるTLPを送るときにこのビットをセットする
Attr：アトリビュート
Last DW BE：ラストDWバイト・イネーブル
1st DW BE：ファーストDWバイト・イネーブル
Compl.Status：コンプリーション・ステータス
BCM（Byte Count Modified）：
PCI-Xデバイスでは，コンプリーションのバイト・カウントに本来の値よりも小さい値をセットすることがあり，そのような場合にBCMビットがセットされる．

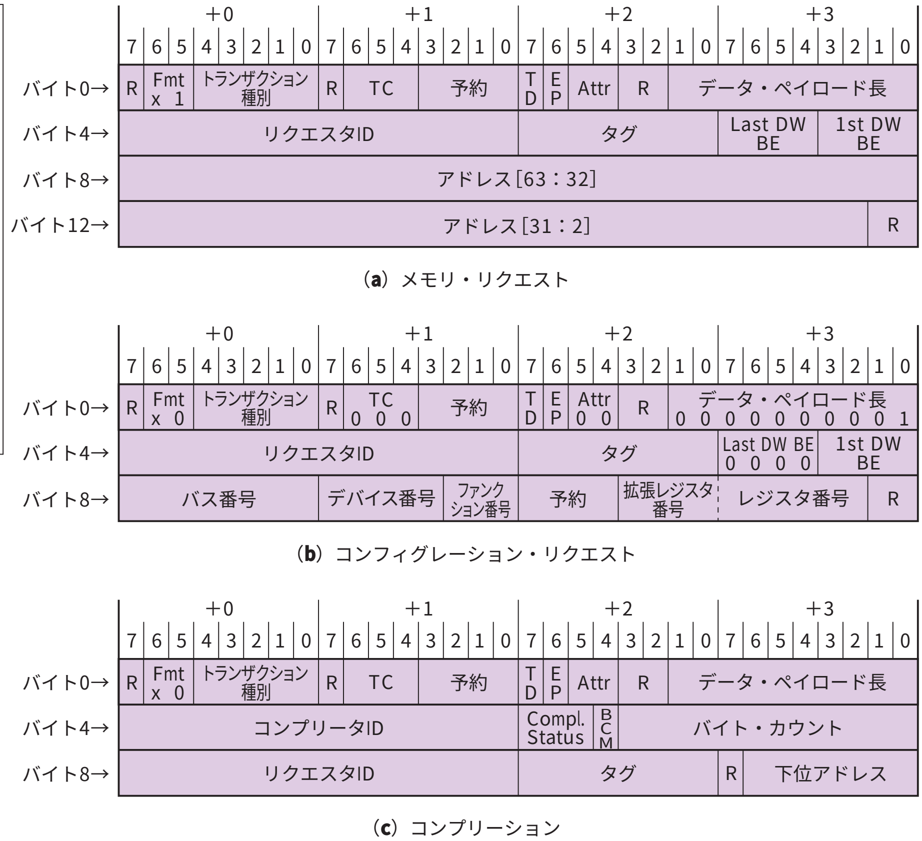


図7
TLPヘッダの例
TLPヘッダの最初の1DW（バイト0～バイト3）は，すべてのトランザクションで共通の仕様になっている．例として，4DW（64ビット・アドレス）のメモリ・リクエスト，3DW（32ビット・アドレス）のコンフィグレーション・リクエストとコンプリーションのTLPヘッダを示す．

NPD（Non Posted Data），CPLH（Completion Header），CPLD（Completion Data）の六つのキュー（待ち行列）に対してフロー制御を行うことになります．

● ビット化けによるエラーを防ぐECRC

PCI Expressにおけるデータ・インテグリティ（データ品質の確保）は，データリンク層のLCRC（Link CRC）で行われますが，オプションでECRC（End-to-end CRC）をTLPダイジェストとしてトランザクション層で付加し，チェックすることもできます．例えば，通信経路にスイッチやブリッジなどのデバイスが入っており，それらのデバイスの内部でビット化けが起きたとしましょう．すると，デバイスからTLPが転送される際に，そのまちがったデータをもとにデータリンク層でLCRCを付加するため，受信側ではエラーを検出できなくなります．

こうした問題を防ぐためにECRCを利用します．ECRCは，データリンク層ではTLPの一部として扱われます．途中経路にあるデバイスによって再計算されることがないので，データの送信元で付加されたCRCがそのまま最終的にデータを受信するデバイスまで届けられます．

● 帯域の制御を行う仮想チャネル

PCI Expressにはオプションとして，一つのポートに仮想的な複数のチャネルを設定し，チャネルごとに帯域を割り当てて帯域制御（QoSやアイソクロナス転送）を行う仮想チャネル（VC：virtual channel）という仕様があります．

しかし，現状では米国Intel社のチップセットでオーディオ・データの転送に使われている程度で，それ以外の用途ではほとんど用いられていません．また，Windows Vistaでもサポートされないため，パソコンやサーバ用途では限定的な使用にとどまるかもしれません．

3.2 データリンク層 ――機器どうしのデータ送受信制御

データリンク層は，トランザクション層から渡されたTLPを確実にリンク先のデバイスへ届けるため，データ・インテグリティという処理を行います．

また，リンク先のデータリンク層と制御情報を受け渡すために，DLLP（Data Link Layer Packet）と呼ばれるデータリンク層専用パケットの送受信も行います．DLLPとしては，以下のものが挙げられます．

- ACK/NAK DLLP：TLPの受信応答および再送要求
- InitFC1/InitFC2，UpdateFC DLLP：トランザクション層で行うフロー制御に必要な情報を送信するためのDLLP
- パワー・マネージメント関連のDLLP

DLLPは，**図8（a）**のように1バイトのDLLP種別，3バイトのDLLPごとの固有情報，そして16ビットCRCの固定長フォーマットです．例えば，ACK DLLPは**図8（b）**のようにバイト2およびバイト3にシーケンス番号が入ります．

● 機器間送受信のためのデータ・インテグリティ

送信側のデータリンク層では，通信途中にパケットの消失が起きたことを検出できるように，すべてのTLPにシーケンス番号と呼ばれる12ビットの一連の番号を付与してTLPといっしょに送信します（**図9**）．また，このシーケンス番号とTLPを合わせたものに対して32ビットのCRCを計算し，それをLCRCとしてパケットの末尾に付加します．

受信側のデータリンク層では，シーケンス番号が直前に受け取ったパケットと続き番号になっているか，CRCエラーがないかどうかをチェックします．チェックの結果，問題がなければ，ACK DLLPを送信側へ返送します．もし問題があれば，NAK DLLPを送ってTLPの再送を要求します．なおACK応答は，データリンク層がTLPを正しく受け取れたことを示すだけで，そのTLPに含まれるコマンドやデータが正しく処理されたかどうかとは無関係です．

ACKおよびNAK DLLPは，TLPごとに返送する必要はなく，複数のTLPに対する応答をまとめて返送することが許可されています．例えば，シーケンス番号n，$n+1$，$n+2$，$n+3$の四つのTLPを受信した場合，ACKを返すときに$n+3$を指定すれば，$n+3$までのTLPを正しく受け取ったことを意味します．あるいは，NAKとして$n+1$を返送すると，$n+1$までのTLPの再送を要求することになります．なお，この場合は送信側は$n+1$のTLPを送った後に$n+2$と$n+3$のTLPから転送を再開します．つまり，NAKで$n+3$を指定した場合と同じです．これは，デバイス内部のバッファ管理をシーケンシャル処理だけで済むように簡易化し，複雑なランダム・アクセスを不要にす

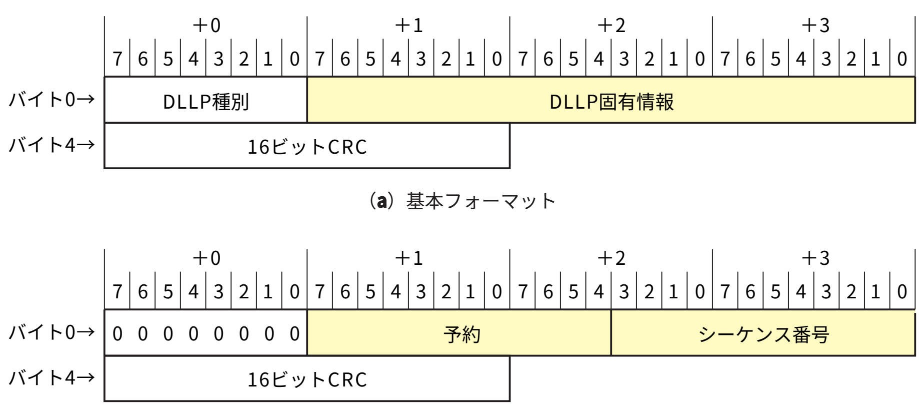


図8
DLLPフォーマット
1バイトのDLLP種別と3バイトのDLLPごとの固有情報，16ビットCRCからなる固定長フォーマット．ACK DLLPの場合，（b）のようにバイト2およびバイト3にシーケンス番号が入る．また，NAK DLLPの場合はDLLP種別に“00010000”が入る．

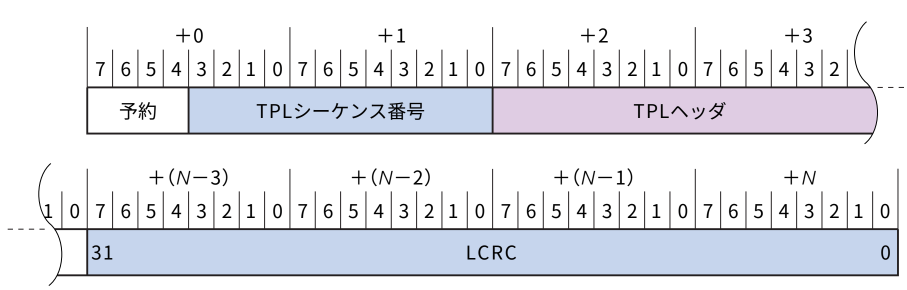


図9
データリンク層におけるTLPの処理
トランザクション層から渡されたTLPを確実にリンク先の機器へ届けるため，データリンク層でTLPの先頭に12ビットのシーケンス番号を付加し，末尾に32ビットのLCRC（Link CRC）を付加してから，送信する．

るためです．

また送信側はTLPを送った後，一定時間たってもACKまたはNAKが返ってこなかった場合もTLPの再送を行います．これらの再送処理は最大4回まで行われます．4回行った後，それでも送信が成功しなかった場合，データリンク層は物理層へリンクの再トレーニングを要求します．

3.3 物理層──符号化処理やシグナリングを行う

物理層は実際にデータを伝送するためのシグナリングや符号化などを行います．物理層の詳細はトランザクション層やデータリンク層からは見えないため，伝送方式の変更に柔軟に対応できます．例えば，現在のGen1（1レーン当たりのデータ転送速度は2.5Gbps）からGen2（同5Gbps）への変更では物理層のみ変えればよく，トランザクション層やデータリンク層のプロトコルに変更はありません．

物理層は，符号化やリンクの制御を行う論理サブブロックとシグナリングを行う電気サブブロックの二つに分けられます（図10）．

● 論理サブブロックの役割

論理サブブロックで処理する内容を以下に述べます．

1）フレーミング

データリンク層から渡されたTLPまたはDLLPへパケットの境界を示すKコード（特殊符号）が付加されます．TLPの場合，先頭にSTP（Start of TLP）が，末尾にENDが付加されます（図11（a））．DLLPの場合は先頭にSDP（Start of DLLP）が，末尾にENDが付加されます（図11（b））．この処理をフレーミングと呼びます．また，TLPの場合は

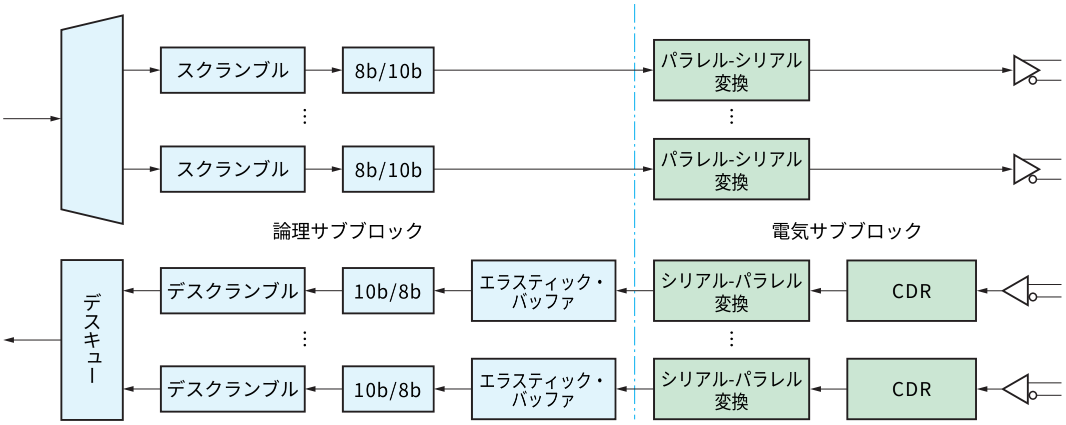


図10 物理層
物理層は，符号化やリンクの制御を行う論理サブブロックとシグナリングを行う電気サブブロックの二つに分けられる．

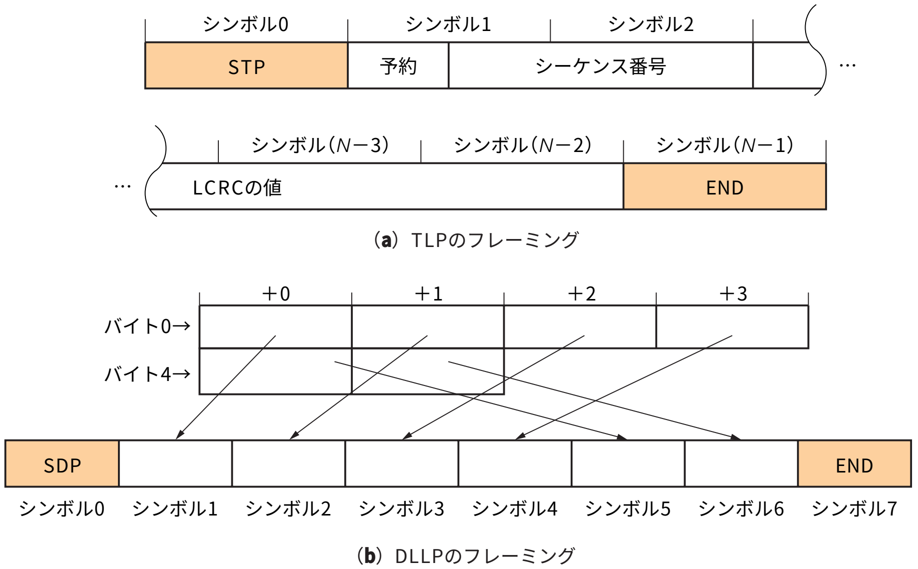


図11 フレーミング
(a)のように，TLPの場合は先頭にSTP，末尾にENDが付加される．また，DLLPの場合は(b)のように先頭にSDP，末尾にENDが付加される．

STPからENDまでのバイト数がDW単位になるように，1～3個のPADと呼ばれるKコードがTLPとENDの間に挿入されることがあります．

なお，以下の説明ではフレーミング後の各バイトのことを便宜上「シンボル」と呼びます（本来，シンボルは8b/10b変換後の10bコードを指す）．

フレーミングが行われた後，リンクのレーン構成に応じて各シンボルがレーンに割り当てられます．x1リンクとx4リンクの例を**図12**に示します．

2）スクランブル

各レーンに分割されたデータは，レーンごとにスクランブルおよび8b/10b変換が行われた後でシリアル・データとして送信されます．

同じデータが連続した場合，そのデータ・パターンに含まれる周波数成分がほかよりも強くなり，EMI（electromagnetic interference）の面から望ましくないため，スクランブルをかけて別々のパターンとして送信を行います．PCI Expressでは，$G(X)=X^{16}+X^5+X^4+X^3+1$を生成多項式とするLFSR（linear feedback shift register）を使ってスクランブルを行います．受信側では同じLFSRを使ってデスクランブルを行います．なお，スクランブル/デスクランブルされるのはデータ・シンボルのみで，Kコードはそのまま送受信されます．

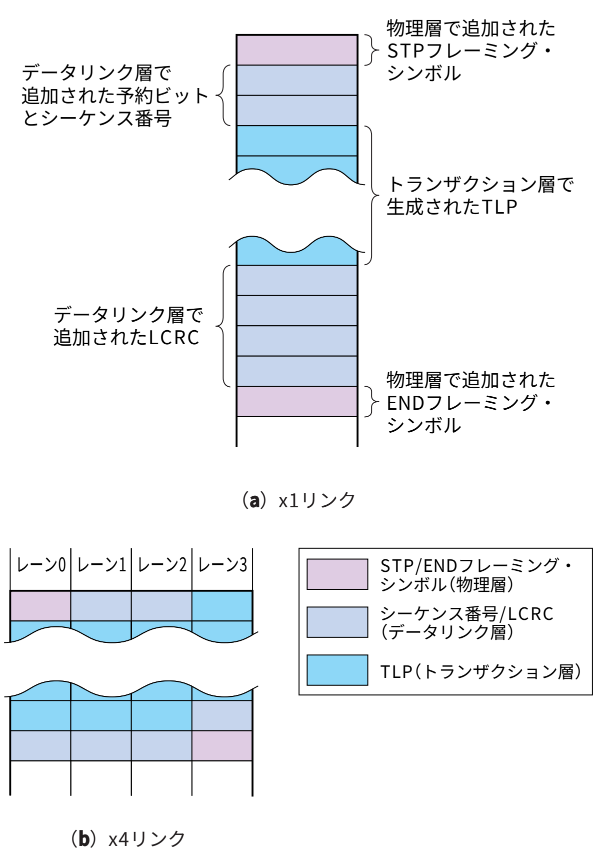


図12　フレーミング後のシンボル

3）8b/10b変換

8b/10bは，8ビットのデータを10ビットに変換して送信するもので，以下の特徴が挙げられます．

- ‘0’や‘1’の連続したデータでも，‘0’と‘1’の変化を多く含むビット・パターンに符号化されるため，受信側でクロック再生が容易に行える．また，それによって信号に低い周波数成分が含まれない
- ‘0’と‘1’の数が同じになるように符号化され，DCバランスがとれ，AC結合が可能
- 符号空間が広がったため，256種類のデータ以外にKコードと呼ばれる制御用の符号が追加される
- 符号に冗長性があるため（1024通りのうち256＋Kコードの数しか使わない），ある程度のエラー検出が可能
- データ転送効率は80％

8b/10bでは，LSB（least significant bit）からMSB（most significant bit）へ向かって8ビットの各ビットをA，B，C，D，E，F，G，Hと呼び，それをEDCBAとHGFの二つのグループに分けます（**図13**）．EDCBA，HGFの順番で，データをD*x.y*，KコードをK*x.y*とします．例えば，データが00110000bの場合はHGF＝001b，EDCBA

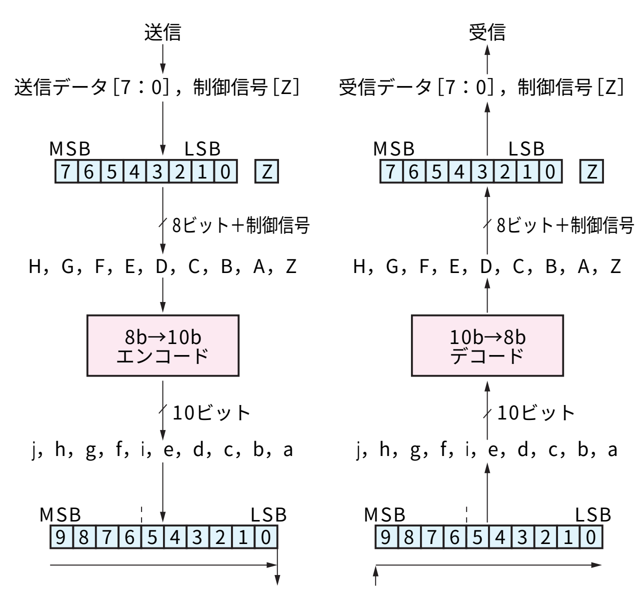


図13　8b/10b変換

図の手順で，送信側から受信側へ8b/10b変換を行う．

＝10000bなのでD16.1となります．このとき，Kコードの01111100bはHGF＝011b，EDCBA＝11100bなのでK28.3となります．

EDCBAを5b/6b変換したものをiedcba，HGFを3b/4b変換したものをjhgfとして，これをシリアルで送信するときはa～jの順番で送信します．

8b/10bにはいくつかのKコードがありますが，PCI Expressでは**表2**に示す9個が使われています．

4）レーン間デスキュー

PCI Expressでは，レーンとレーンの間のスキューは受信側の回路で合わせ込みを行います．例えば，**図14**のように各レーンにFIFO（first-in first-out）メモリを設けて，スキューを吸収する方法などが一般に用いられています．

これによって，ボード上ではレーン間の配線長の差をほとんど気にせずに配線を行えます．

5）エラスティック・バッファ

リンクの両端のデバイスで使われるリファレンス・クロックには，最大±300ppmの周波数偏差が許容されています．このため，受信側よりも送信側のクロックのほうが速い，もしくはその逆が起こりえます．この周波数偏差による送信側と受信側の間の速度差を吸収するため，エラスティック・バッファと呼ばれるFIFOメモリが用いられます．

送信側は最低1,180シンボル，最高1,538シンボルごとに1回，COMシンボル（**表2**を参照）の後に3個のSKPシンボルが続くSKIPオーダード・セットを送信データに挿入します．受信側ではSKIPオーダード・セットを含むすべての受信データをエラスティック・バッファにいったん格納します．そして，COMシンボルの後に続くSKPシンボルの数を1～5個に調整することによって，送信側と受信側の間の速度差を調整します．送信側のほうが速いときは受信側はSKPシンボルを間引き，逆に送信側が遅いときは余分にSKPシンボルを挿入することで速度差を吸収しているわけです（**図15**）．

表2　PCI Expressに用いるKコード

Kコード	シンボル	内　容
K28.5	COM	レーン/リンクの初期化および制御に使う
K27.7	STP	TLPの始まりを示す
K28.2	SDP	DLLPの始まりを示す
K29.2	END	TLP/DLLPの終わりを示す
K30.7	EDB	送信途中でエラーが発生したTLPでENDの代わりに用いる
K23.7	PAD	フレーミング時のパッド・キャラクタ，およびリンク初期化の際に用いる
K28.0	SKP	二つのポート間での周波数偏差を補償するために挿抜される
K28.1	FTS	L0sからL0への遷移時のトレーニング・シーケンスに用いる
K28.3	IDL	EIオーダード・セットで用いる

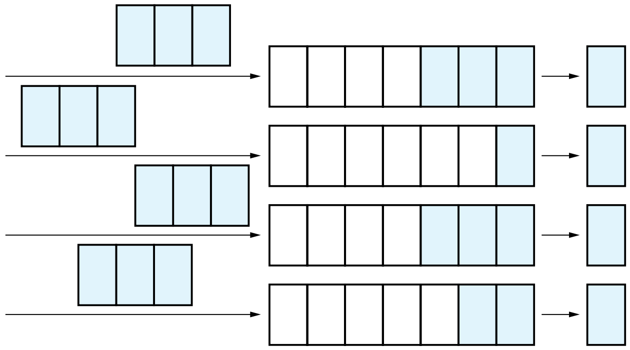


図14　デスキュー
図のように各レーンにFIFOメモリを設けて，スキューを吸収する方法が一般に用いられている．

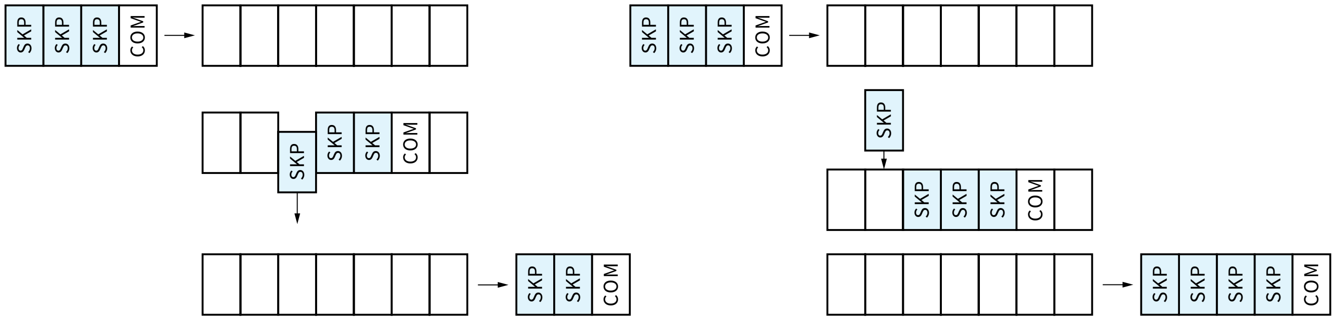


図15　エラスティック・バッファ
送信側のほうが速いときには，（a）のように受信側はSKPシンボルを間引き，逆に送信側が遅い場合は，（b）のように余分にSKPシンボルを挿入することで速度差を吸収している．

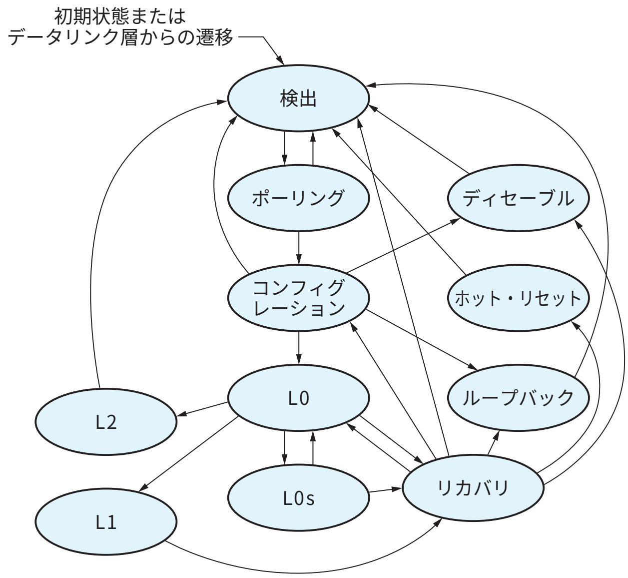


図16　LTSSM
LTSSMのステート・マシンを示す．実際はサブステートが存在し，もう少し複雑になる．

6）LTSSM（Link Training and Status State Machine）

LTSSM（Link Training and Status State Machine）は，リンクの初期化やコンフィグレーションを制御するもので，図16に示すようなステート・マシンで制御されます．なお，実際には図の各状態にさらにいくつかのサブステートが存在し，比較的複雑なステート・マシンになっています．

検出（detect）状態では対向デバイスが存在するかどうかをチェックし，存在すればリンクの初期化を開始します．

ポーリング（polling）状態では，対向デバイスの使用可能なレーンの数やビット・レートをチェックします．また，対向デバイスが送ってきたTS1オーダード・セットやTS2オーダード・セットを使って，受信側のビット・ロック，シンボル・ロックも同時に行われます．さらに，PCI Expressではボード上の配線を行いやすくするため，差動信号の＋と－を逆につなぐことを許されているので，極性が反転したTS1/TS2を受け取ることも考えられます．そのような状態を見つけた場合，受信側が内部で極性反転を行います．

コンフィグレーション（configuration）状態では，ポーリング状態で検出したレーンを使用できるように設定します．複数レーンの場合，自分のレーン番号の並びとは逆の並びを相手側のデバイスが送ってくることがあります．これは極性反転の場合と同じく，ボード上の配線を容易にするために同じ番号のレーンどうしを接続しなくてもよいためです．アップストリーム・ポートがオプションのレーン反転をサポートしている場合，レーン番号を逆順に変えてリンクを確立します．また，レーン間デスキューはコンフィグレーション状態の終わりまでに完了しなければなりません．

リンクのすべてのコンフィグレーションが終わると，LTSSMはL0へ遷移します．L0は二つのデバイス間でリンクが確立した状態であり，TLPの送受信はL0のときのみ行われます．

LTSSMには上記以外に，パワー・マネージメントで使われるL0s状態やL1状態，リンクが一時的に不調になったときに復旧するためのリカバリ（recovery）状態などもあります．

7）オーダード・セット

物理層ではTLPやDLLP以外にも，オーダード・セットと呼ばれる物理層のリンク制御に用いられる一連のシンボルが送られます．オーダード・セットには，リンク・トレーニングに使われるTS1/TS2オーダード・セット，リンク両端のデバイス間の周波数偏差を吸収するためのSKIPオーダード・セット，L0sからL0へ復帰する際のリンク・トレーニングに使われるFTSオーダード・セット，そしてリンクが電気的アイドル状態に入ることを通知するEIオーダード・セットがあります．

● 電気サブブロックの役割

電気サブブロックは，実際にシリアル通信を行うトランスミッタ（送信回路）やレシーバ（受信回路）から構成されます．レシーバには，入力されたビット・ストリームからクロックとデータを分離・再生するCDR（clock data recovery）と呼ばれる回路が含まれます．

PCI Expressなどに使われる伝送路では，周波数の高い信号，つまり‘0’と‘1’の変化の多い信号ほど減衰が大きく，周波数の低い信号，つまり‘0’と‘1’の変化の少ない信号は減衰が少ないという特性を示します．そのため，減衰量の差をあらかじめ考慮して‘0’や‘1’が連続する場合，はじめのビットよりも出力振幅を少なくして，レシーバで見たときに同じような振幅になるように補正を施します．これを「デエンファシス」と呼びます．PCI Expressでは，デエンファシスは3.5dBと規定されています．

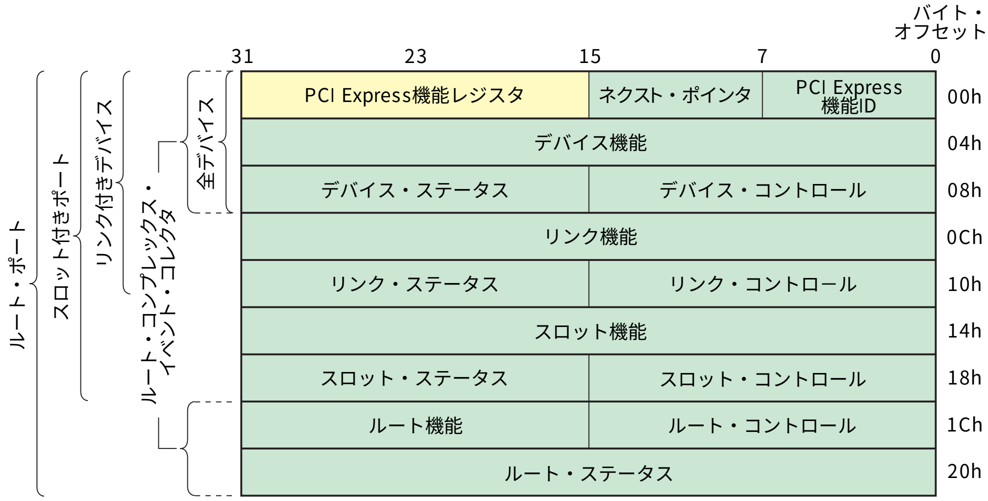


図17
レジスタ構成
PCI Expressデバイスは，PCI Express機能レジスタと呼ばれるレジスタを互換領域にかならず持つ．

3.4 ソフトウェア層——従来PCIとの互換性を維持

トランザクション層の上位にはソフトウェア層があります．PCI Expressに追加された機能を用いなければ，基本的に従来のPCI/PCI-Xのソフトウェアをそのまま利用できますが，そうでない場合はなんらかの変更が必要になります．

以下に，PCI Expressのソフトウェア層にかかわる仕様について説明します．

● コンフィグレーション・レジスタは4Kバイトに

PCI Expressではコンフィグレーション空間が拡張されました．これに伴い，従来は256バイトだったレジスタ空間が4,096バイトに拡張されました．拡張されたレジスタ空間において，先頭の256バイトの領域を「互換領域」，残りの領域を「拡張領域」と呼びます．

従来のCF8h/CFChのI/Oポートを使ったコンフィグレーション・アクセス方法では，互換領域にのみアクセス可能です．拡張領域も含めてアクセスするためには，PCI Expressで新しく定義されたメモリ・マップド・コンフィグレーション・アクセスを用いる必要があります．このためのメモリ空間をどこへアサインするかは，システムやチップセットの仕様で自由に決められます．OSやドライバはPCIファームウェア仕様で定められたAPI(application programming interface)を使って，このメモリ空間のアドレスを取得します．

PCI Expressデバイスは“PCI Express機能レジスタ(Capability Registers)”を互換領域にかならず備えられなければなりません(図17)．

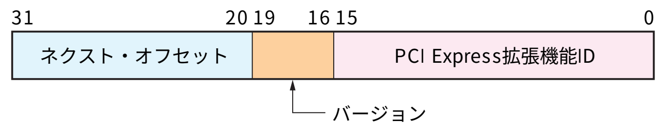


図18　拡張PCI Express機能レジスタのヘッダ
リンクのいちばんはじめのレジスタは100hに置かれ，そこからはネクスト・オフセットで次のレジスタへリンクされる．

一方，オプションとして拡張PCI Express機能レジスタと呼ばれるものもあります．これらは拡張領域に置かれ，各レジスタは図18に示すようなヘッダでかならず始まらなければなりません．リンクのいちばんはじめのレジスタは100hに置かれます．そこからはネクスト・オフセットで次のレジスタへリンクされます．拡張PCI Express機能レジスタには，詳細なエラー情報を格納するアドバンスト・エラー・レポート機能や，仮想チャネルを制御するための仮想チャネル機能などがあります．

参考・引用*文献

(1)* PCI-SIG；PCI Express Base Specification Revision 1.1, March 28, 2005.

のざき・はじめ
米国NEC Electronics America社

<筆者プロフィール>
野崎原生．1988年，NEC入社．2003年より米国NEC Electronics America社に勤務．現在，PCI ExpressやAMBなどの高速シリアル・インターフェースを備えた製品の開発を行っている．いつまでたっても英語が上達しないことが目下の悩み．

2